김소연, 김다윤, 강우원

NLP 감정분석 기반
마케팅 시장 분석

온라인 커뮤니티 특화 감성 사전 구축을 위한 새로운 용어 극성값 분석 시스템

빅데이터미네이터

“ ☹😐☺

주제 :
온라인 커뮤니티 특화 감성 사전 구축을 위한 새로운 용어 극성값 분석 시스템 ”

지난주 주요 피드백

→ 새로운 용어가 기존 사전의 어떤 용어와 비슷한지 찾는 기법 구현에 대한 고민 필요

→ 특허에 출원할 수 있는 아이디어

- 1:1, 1:n 매핑 : 특정 새로운 단어가 사전과 매핑될 수 있도록

- 시간의 축 : 단어의 뜻이 시간에 따라 변경 되는 부분 분석

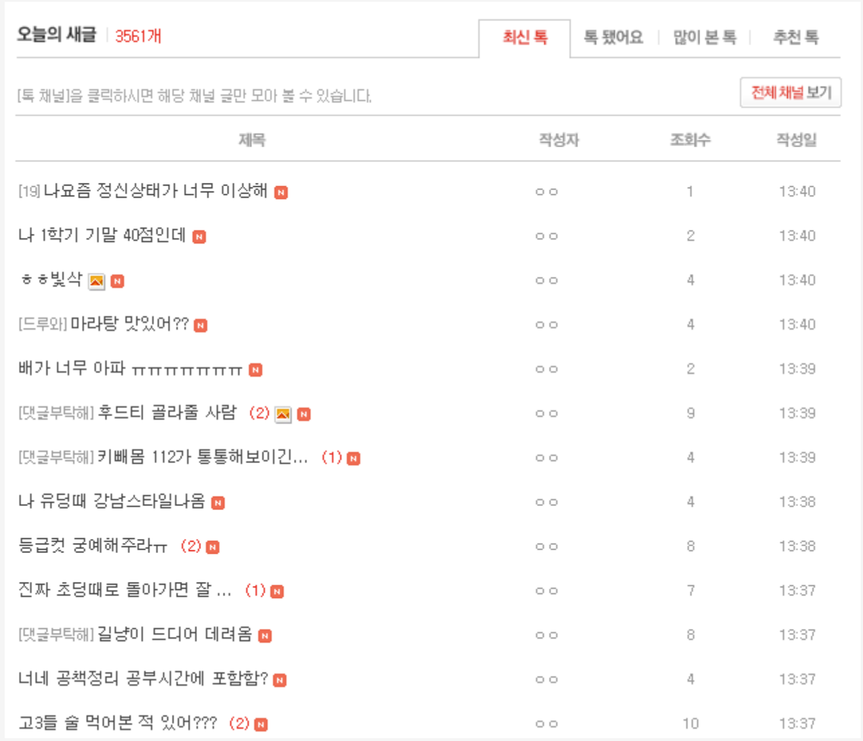
구현 계획 변경 사항

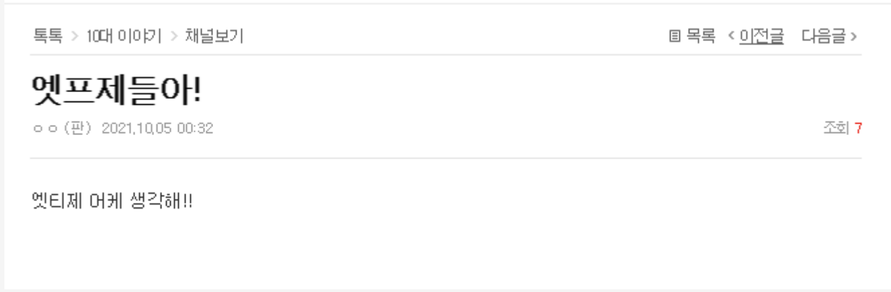
- “네이트판”으로 커뮤니티 변경

  → 크롤링을 진행하면 “too much results” 에러가 발생하는 FM코리아의 문제점

  → FM코리아는 우회를 시도해도 우회 방지로 크롤링을 차단

  → 네이트판은 10대 사용률이 높아 신조어가 가장 많이 출현해서 새로운 용어 추출에 용이함





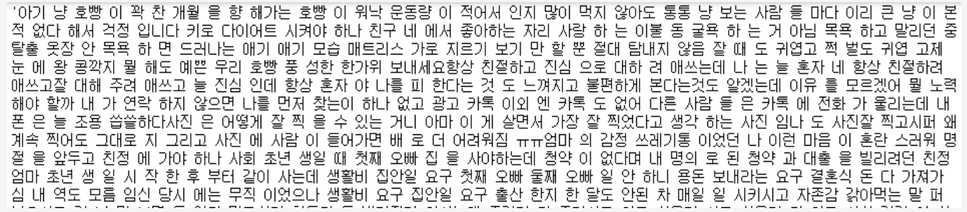
진행 사항: 새로운 용어 추출 및 분석

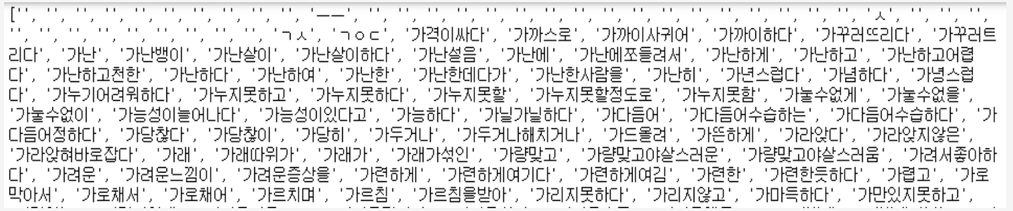
- KNU 감성 사전에 없는 단어 비교 추출

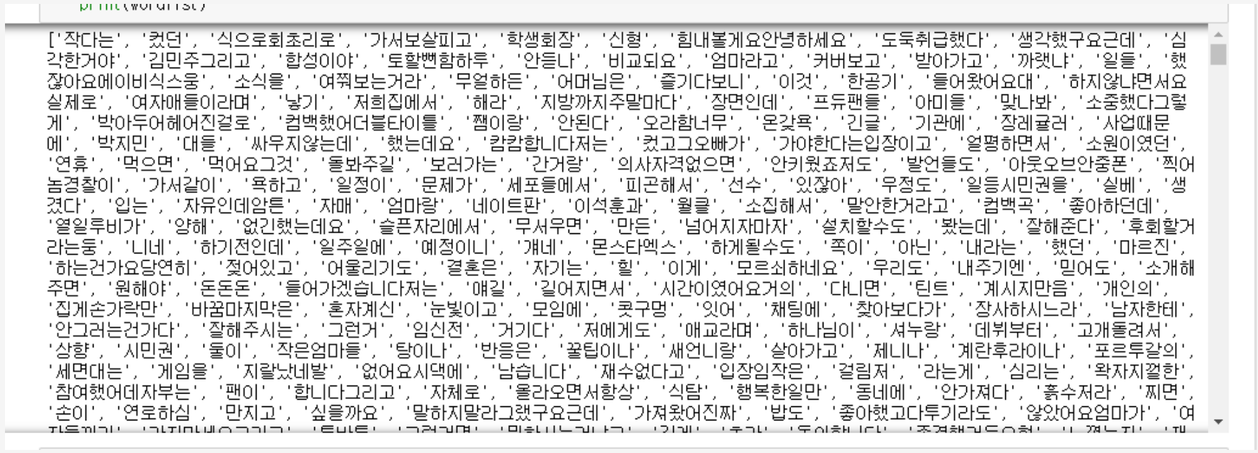
1. 크롤링 데이터 추출

2. KNU 감성 사전 데이터 추출

3. 두개를 비교하고 새로운 단어 추출해서 txt로 저장







진행 사항: 새로운 용어 추출 및 분석

- Soynlp 단어 추출

1. Soynlp 패키지를 사용해서 비지도 학습 기반의 명사 추출
2. 추출한 명사 중 새로운 단어 판별

```
from soynlp import DoublespaceLineCorpus
```

```
# 문서 단위 말뭉치 생성
corpus = DoublespaceLineCorpus("natepann.txt")
len(corpus)  # 문서의 갯수
```

```
251
```

```
# 문장 단위 말뭉치 생성
corpus = DoublespaceLineCorpus("natepann.txt", iter_sent=True)
len(corpus)  # 문장의 갯수
```

```
451
```

```
from soynlp.word import WordExtractor

word_extractor = WordExtractor()
word_extractor.train(corpus)
```

```
training was done. used memory 0.103 Gby 0.092 Gb
```

```
word_score = word_extractor.extract()
```

```
all cohesion probabilities was computed. # words = 1353
all branching entropies was computed # words = 2561
all accessor variety was computed # words = 2561
```

진행 사항: 새로운 용어 추출 및 분석

- Kiwipy 단어 추출

1. 전처리 후 사전 미등록 명사 추출해주는 kiwipy 기능 사용
2. 새로운 용어를 명사 단위로 추출, 사용자 사전에 추가

```
In [15]: inputs = list(open('natepann.txt', encoding='utf-8'))
         kiwi.extract_words(inputs, min_cnt=1, max_word_len=10, min_score=0.2, pos_score=-3.0, lm_filter=True)

Out[15]: [('부모님', 1.174511194229126, 47, -1.0527305603027344),
          ('예비신랑', 0.9363517761230469, 20, -1.2639905214309692),
          ('스트레이키즈', 0.8020613193511963, 5, -2.1006152629852295),
          ('더보이즈', 0.6040971279144287, 8, -2.790191650390625),
          ('대왕트래블', 0.5548834204673767, 8, -0.7859810590744019),
          ('엔하이픈', 0.4540952146053314, 6, -2.910804271697998),
          ('엔시티', 0.4487341642379761, 10, -2.095034599304199),
          ('다대왕트래블', 0.3298221826553345, 5, -1.5150139331817627),
          ('시부모님', 0.32935193181037903, 12, -0.8014914989471436),
          ('임신출산육아', 0.2775214612483978, 3, -2.0598337650299072),
          ('예비신부', 0.25270774960517883, 7, -1.7872761487960815)]
```

→ 사용자 사전 추가 가능해서 사전에 등록 안된 명사들 추가 필요

→ 5일치 데이터라 미등록 명사 추출 양이 적었음, 데이터 양을 늘릴 필요

→ kiwipy 방식이 가장 정확성이 높았음

진행 사항: 새로운 용어 분석 아이디어

- Topic 분류 후 새로운 용어와 기존 단어를 매핑

1. 게시글을 보면 글마다 topic 분류가 되어있음
2. 이를 이용해서 데이터마다 topic을 분류해주고 레이블링하는 방식


```
시부모      -1, "결혼/시집/친정"
시부모      +1, "요리&레시피"
시부모       0, "이슈"
```

→ 단어마다 분류된 topic을 확인하고 분석에 사용하면, 글마다 단어가 다른 의미로 사용되는 문제를 해결할 수 있을 것으로 생각된다. Ex) 결혼/시집/친정 topic일때 '부정 ' 인 단어가 사는 얘기 topic에서 '긍정 ' 인 단어로 사용되는 경우

→ 새로운 단어가 들어오면 단어의 topic을 먼저 본 후, 단어의 연관 단어를 그 topic으로 분류된 단어만 추출해서 새로운 용어를 기존의 단어와 좀 더 정확하게 매핑시킬수 있도록 한다.

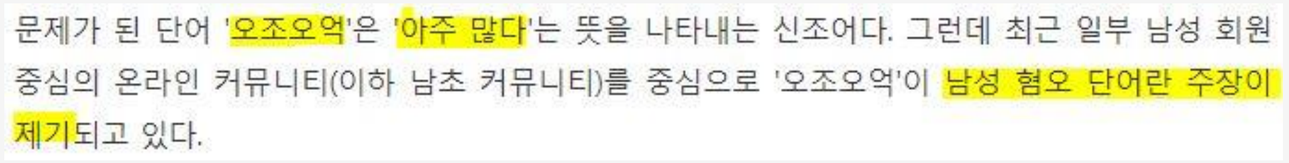
진행 사항: 새로운 용어 분석 아이디어

- 사용자의 특성에 따라서 달라지는 새로운 용어 의미 파악

  - 사용자의 나이, 성별, 커뮤니티 특성

→ 새로운 용어 입력시 미리 특성에 따라 분류된 커뮤니티에서 해당 단어를 검색하여 그 극성을 파악하고 어떻게 사용자에 따라 긍부정 극성이 달라지는지 수치화와 해당 단어가 사용되는 예시 문장을 제공

Ex) 멘토님 의견(요즘 사용되는 단어들)


문제가 된 단어 '오조오억'은 '아주 많다'는 뜻을 나타내는 신조어다. 그런데 최근 일부 남성 회원 중심의 온라인 커뮤니티(이하 남초 커뮤니티)를 중심으로 '오조오억'이 남성 혐오 단어란 주장이 제기되고 있다.

차주 계획

1. 크롤링 데이터를 늘려서 kiwipy 사용하여 새로운 용어 추출
2. 논문 작성 및 제출 - 한국멀티미디어학회

감사합니다 :)

빅데이터미네이터

손윤식 교수님 주식회사 닥터송 송환구 멘토님